灰姑娘能变成白雪公主吗？感知到的信任对他人面孔表征的影响*

李庆功[1] 方溦[2,3] 胡超[4] 石德君[5]

胡晓晴[6] 傅根跃[7] 王乾东[8]

（[1]浙江师范大学心理学院，浙江省儿童青少年心理健康与危机干预智能实验室，金华 321004）

（[2]浙江理工大学心理系，杭州 310018）

（[3]麦克马斯特大学心理系，汉密尔顿 L8S 4L8）

（[4]东南大学人文学院医学人文系，南京 211189）

（[5]北京大学心理与认知科学学院，北京 100871）

（[6]香港大学心理学系，香港大学脑与认知科学国家重点实验室，香港 999077）

（[7]杭州师范大学经亨颐教育学院心理系，浙江省哲学社会科学培育实验室"杭州师范大学婴幼儿发展与托育实验室"，浙江省认知障碍评估技术研究重点实验室, 杭州 311121）

（[8]北京师范大学心理学部，应用实验心理北京市重点实验室，心理学国家级实验教学示范中心，北京 100875）

摘要 本研究考察对他人可信程度的感知是否会影响对该人物面孔长相的表征及其潜在的机制。实验 1 让被试形成目标人物可信或不可信的印象。随后利用反相关图像分类技术将被试对目标人物面孔的心理表征可视化。结果发现无论目标人物是男性还是女性，高可信度的目标人物与更具吸引力和积极特质的面孔表征相关。实验 2 从一批新的被试中可视化了可信和不可信群体的面孔表征的特征，并与实验 1 中获得的目标人物的面孔表征的特征做相似性分析，发现被描述为可信(或不可信)的目标人物的面孔表征特征与可信(或不可信)群体的面孔表征特征有更多的相似性，说明当人们得知他人是可信(或不可信)时，会把脑海中的对应图式特征叠加到该人物的面孔物理特征上，从而重塑面孔表征。本研究说明自上而下的

收稿日期：2022-12-19
通信作者：王乾东，E-mail: wangqd@bnu.edu.cn
作者简介：李庆功和方溦为共同第一作者。
* 中央高校基本科研业务费专项资金（2021NTSS60）；国家自然科学基金项目（32200872，62176248，62176248）；教育部人文社科青年项目（22YJC190022）资助。

加工方式在面孔表征形成中扮演了重要作用。

关键词 人际知觉，反相关图像分类技术，心理表征，吸引力，可信度

分类号 B842

1 引言

人们往往从面孔长相中推断一个人的特质(Bonnefon et al., 2017; Lin et al., 2021; Sutherland & Young, 2022; Todorov et al., 2015)，例如判断他人的可信度和能力水平(Todorov et al., 2009)，并对个体做出积极(或消极)的社会评价(Bascandziev & Harris, 2014; Chen et al., 2014)。研究发现人们普遍认为相貌好看的个体会具备更多优秀的品质，对其产生“美即是善”的刻板印象(Dion et al., 1972)。此外，对个体品质的了解也可以在一定程度上改变人们对于该个体面孔吸引力的感知(Bliss-Moreau et al., 2008; Todorov & Uleman, 2002, 2003, 2004)，一个典型的情形是“情人眼里出西施”，个体倾向于对自己喜欢的面孔做出更正性的评价，并且认为诚实的人比不诚实的人更具有外表吸引力(Paunonen, 2006)。

然而，对个体特质的知觉是否影响对该人物面孔的心理表征，以及如果影响，其背后的机制尚不明晰。心理表征是指外部事物在心理活动中的内部再现，它一方面反映客观事物，另一方面又是心理活动进一步加工的对象。由于其可能成为心理活动加工的对象，对个体面孔的心理表征可能会随着对该人物态度的改变而改变。根据“眼见为实”的说法，对个体的面孔表征体现出一种“自下而上”的视觉加工方式，人们对个体的面孔表征应该仅仅反映该人物的面孔物理特征，而不会受到印象这种 “自上而下”的加工方式的影响。然而，有研究表明人们对于社会群体的面孔表征会受到“自上而下”的刻板印象的影响(Dotsch et al., 2008; Lloyd et al., 2020)。比如，偏见程度较高的被试会将摩洛哥人的面孔表征为更有犯罪倾向且更不可信(Dotsch et al., 2008)。本研究探究当人们对个体产生不同的印象时，对其所形成的心理面孔表征是否也会不同。已有研究表明，对非裔美国人具有更深肤色且更符合其传统刻板印象的面孔表征的人会对非裔美国人做更加不公平的行为，如分配更少的资金(Krosch & Amodio, 2014)。因此，对面孔的心理表征会直接影响个体的决策行为(Ratner et al., 2014)。本研究从心理面孔形成的角度解析面孔加工的特点和机制，对于进一步了解个体的认知特点与行为方式有重要意义。

本研究旨在考察个体的可信程度是否会影响表征该个体的心理面孔及其心理面孔形成的潜在机制。实验 1 主要探究的问题是：同一张面孔在被描述为可信时，该面孔的心理表征是否会比被描述为不可信时更具有吸引力？如果两种可信程度条件下被试表征出来的面孔存在差异，则可以认为自上而下的特质知觉会影响对他人心理面孔的表征(两组条件下的被试看到的是一模一样的面孔，消除了自下而上的面孔物理特征的影响)。本研究特别关注面孔吸引力与可信度感知之间的关系，是因为已有研究证实，在同时评估这两个特征时，二者

之间存在高度相关(Dion, 1972; Langlois et al., 2000; Mende-Siedlecki et al., 2013)。同时，吸引力和可信度的知觉是社会感知和社会互动的重要成分。

一般来说，心理面孔的表征比较抽象，并且丰富的面孔特征可能很难进行有意识的内省或自我报告。反相关图像分类(reverse correlation image classification, RCIC)技术能够将心理表征的内容直观可视化，揭示人们的内在表征和决策策略(Dotsch & Todorov, 2012; Gosselin & Schyns, 2003)。该技术是数据驱动的，需要被试在多轮试次(往往至少 300 试次)中从两张模糊的面孔(同一张面孔底片叠加不同的随机视觉噪音)中选择最能代表目标个体或特定社会群体的面孔。通过在许多试次中做出这些选择，被试本质上提供了与他们的心理表征相关的图像特征的信息。对被试选择的模糊面孔的噪音叠加平均后与面孔底片合成即获得了被试的分类图像。由于该图像显示了驱动感兴趣的社会判断的刺激特征，因此被视为内在的心理表征。本研究中的心理面孔表征的操作化定义即为使用 RCIC 技术获得的分类图像作为可视化的心理面孔表征。反相关图像分类技术被广泛应用于社会知觉相关的研究中，以调查不同社会类别的成员在感知者的头脑中是如何被表征的。例如，研究者已经探究了人们如何表征外族或者外群体成员(Dotsch et al., 2008; Dotsch et al., 2011)、亲密伴侣(Karremans et al., 2011)、总统候选人(Young et al., 2014)、警察(Lloyd et al., 2020)，甚至自我面孔(Maister et al., 2021; Moon et al., 2020)等。利用该技术，已有研究者获得了被试对可信和不可信群体的面孔表征；随后，另一批被试对这两张面孔表征进行评价，发现相比于不可信群体的面孔表征，被试评价可信群体的面孔表征更值得信赖(Dotsch & Todorov, 2012)。本研究在 Dotsch 和 Todorov (2012)基础上，使用反相关图像分类技术考察在印象形成任务中，操纵目标人物的可信度对该人物面孔表征的影响。我们假设当目标人物被描述为可信时，被试表征的目标人物的面孔会更具吸引力。

实验 2 进一步揭示个体感知到的信任水平影响表征他人面孔的潜在心理机制。实验假设个体在形成他人面孔的心理表征时，需要将该面孔的物理特征(即“自下而上”的加工方式)与该面孔所属的社会群体的刻板特征(即“自上而下”的加工方式)相结合。因此，当个体知道他人值得信任时，个体可能会对目标人物的面孔进行二次加工，即将其所认为值得信赖的个体所属的群体特征(可信群体面孔表征特征)赋予该目标面孔。如果该假设是正确的，则可以预期可信/不可信的目标人物的面孔表征特征与可信/不可信群体的面孔表征特征有更多的相似性。实验 2 使用 Dotsch 和 Todorov (2012)中描述的标准反相关图像分类实验程序获得可信和不可信群体的面孔表征特征，并与实验 1 获得的面孔表征特征进行相似性分析。

2 实验 1

实验 1 探究对女性和男性目标个体面孔的心理表征是否会受到目标个体的可信程度的影响。实验分为两个阶段。在第一阶段，被试需要先完成印象形成任务，即对目标人物形成可信或者不可信的印象。随后，被试完成反相关图像分类任务，他们需要回忆并从一系列两张叠加噪音的面孔中选出更像该目标人物的面孔，产生面孔分类图像(classification images，CIs)，以此可视化对于目标个体面孔的心理表征。第二阶段会招募另一批被试对第一阶段的分类图像进行特质评估，主要关注被描述为可信人物所产生的分类图像是否会被评价为更具吸引力。

2.1 研究方法

2.1.1 被试

在第一阶段中，我们总共招募了 155 名被试，随机分配到四个组中(2 目标面孔性别：男和女 × 2 可信水平：可信和不可信)。其中，男性目标面孔可信组有 40 名被试(20 名女被试)，平均年龄为 20.38 岁，标准差 1.85 岁；男性目标面孔不可信组有 40 名被试(20 名女被试)，平均年龄为 19.68 岁，标准差 1.82 岁；女性目标面孔可信组有 37 名被试(20 名女被试)，平均年龄为 19.86 岁，标准差 1.60 岁；女性目标面孔不可信组有 38 名被试(21 名女被试)，平均年龄为 20.42 岁，标准差 1.95 岁。

第二阶段有两个任务，该阶段招募的所有被试均未参加第一阶段的实验。任务 1 要求 40 名被试(女性 20 人，平均年龄为 19.98 岁，标准差 1.29 岁)评估第一阶段产生的分类图像的吸引力。任务 2 重复并拓展了任务 1 的结果，重新招募了 50 名被试(女性 27 人，平均年龄为 21.32 岁，标准差 2.51 岁)，让他们不仅评价面孔吸引力，还额外评价多种其他特质(包括可信、聪明、友善、积极表情、刻薄、贪婪、攻击性和支配性)。

所有被试均为在校大学生，视力或矫正视力正常。研究获得了浙江师范大学的伦理委员会的批准。被试在参与实验前签署书面知情同意书，实验后获得一定的金钱报酬。

2.1.2 实验材料

目标人物面孔 共 40 名被试(20 名女性)(未参与任何正式实验任务)观看 80 张面孔图片并对面孔吸引力(1 = 毫无吸引力，9 = 非常有吸引力)进行打分，从中选出 4 张具有中等水平吸引力的面孔图片作为目标人物面孔(男、女各一张)和干扰人物面孔(男、女各一张)。男女目标人物和男女干扰人物的平均吸引力评分分别为 5.03、5.18、5.13、4.90，与吸引力中间值 5 做比较的单样本 t 检验结果分别为 $t(39) = 0.11, p = 0.910$, $t(39) = 0.76, p = 0.455$, $t(39) =$

0.47, p = 0.644, t(39) = -0.41, p = 0.682。目标人物面孔是被试需要观察和再认的面孔(见实验流程)。选择吸引力中等的面孔，是为了避免天花板或地板效应。

面孔底片 将目标人物面孔与同性别的干扰人物面孔混合(morph)，生成拥有目标和干扰人物面孔各 50%物理特征的面孔(男、女各一张)，如图 1 所示。



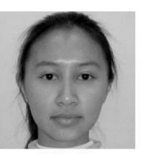






图 1 以女性面孔为例展示了实验 1 中的目标人物面孔、干扰人物面孔和面孔底片(第一行)；第二行展示了面孔底片叠加正(或负)的随机视觉噪音所合成的图片

反相关图像分类任务中的刺激 在面孔底片上叠加随机的视觉噪音而产生。噪音由 6 个不同方位/朝向(orientation) (0°、30°、60°、90°、120°和 150°) × 5 种空间频率(1、2、4、8 和 16 周期/图) × 2 种相位(0 和 π/2) × 随机波幅组成(Dotsch et al., 2008)。共生成 640 对面孔刺激图片，每对中的两张面孔刺激图片叠加的是正负相反的噪音(图 1)。本研究使用的 640 对刺激远多于以往相似的实验范式(如 Dotsch & Todorov, 2012)，采用相反的随机噪音模式也能将所呈现的两张刺激图片间的差异最大化，提升视觉对比效果，从而减少不必要的实验试次数(Dotsch & Todorov, 2012)。每张图片大小为 512 × 512 像素。

对可信目标人物的描述 “小李，女(男)，20 岁，大学在读。她(他)的室友说她(他)非常值得信任，在借用室友的东西之前总是会征得他人的允许。有一次，小李捡到一个装有巨额现金的钱包，把它还给了失主。”

对不可信目标人物的描述 “小李，女(男)，20 岁，大学在读。她(他)的室友说她(他)非常不值得信任，在借用室友的东西之前从不征求他人的允许。有一次，小李捡到一个装有巨额现金的钱包，没有把它还给失主。”

2.1.3 实验流程和数据处理

图 2 展示了实验 1 的具体流程。

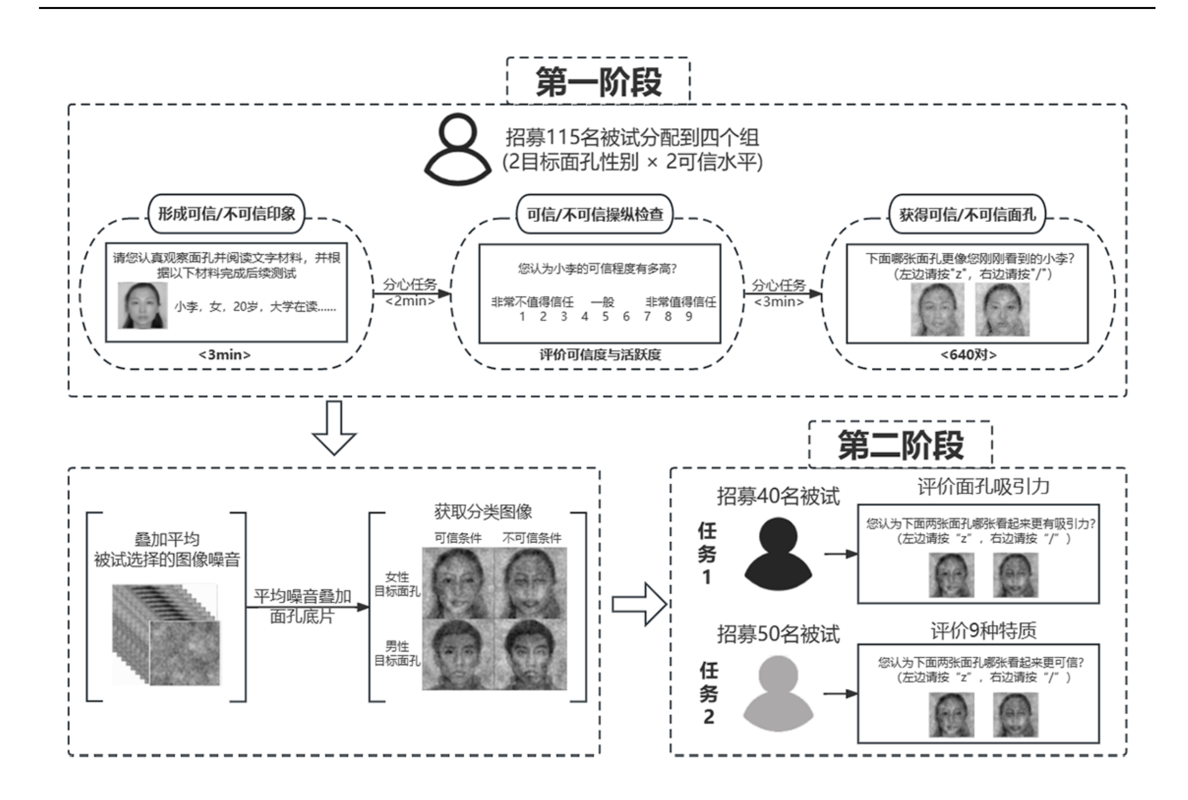


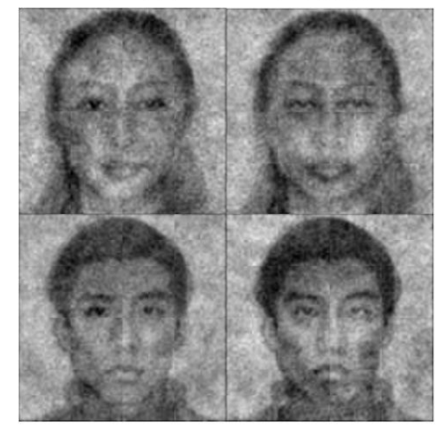
图 2 实验 1 的流程示意图

第一阶段 获取分类图像

被试被要求在三分钟内观察一张男性或者女性目标人物的面孔以及阅读对该人物的简要描述。目标人物在可信条件下被描述为是值得信任的，在不可信条件下被描述为不值得信任(见材料中可信/不可信的目标人物的描述)。被试被告知需要认真观察面孔和阅读文字材料，并被告知在之后的任务中会有一些相关测试。在被试完成 2 分钟的分心任务(从 999 开始倒数并把结果写在纸上)后，要求被试对目标人物的可信度(1 = 非常不值得信任，9 = 非常值得信任)进行评分，以确保信任操作有效。此外，为了确保信任操纵不影响与信任无关的特质，我们也让被试对目标人物的活跃度(1 = 非常不活跃，9 = 非常活跃)进行评分。选择活跃度作为对照指标也可以检验另一个自变量(目标面孔性别)所引起的特有心理效应：我们预期大众会觉得男性比女性更活跃。

在完成另一个 3 分钟的分心任务后，被试进入反相关图像分类阶段，以可视化其对目标人物面孔的心理表征(Dotsch & Todorov, 2012)。每个试次将并排呈现两张模糊的面孔(底片叠加正负相反的随机噪音)(见图 1 和图 2)。被试需要从两张面孔中选择一张最像在第一阶段(即印象形成阶段)中看到的目标面孔。总共有 640 个试次，每个试次的噪音模式都是随机产生的。

对被试选择的 640 张图像的噪音进行平均获得每名被试的噪音模式，再对各个实验条件下所有被试的噪音模式平均，并叠加在原始的面孔底片上，即产生了图 3 展示的在可信和不可信条件下，女性及男性目标面孔的分类图像(即面孔表征)。



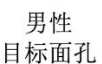


图 3 可信和不可信条件下，女性和男性目标面孔的分类图像(所有被试的平均)

为了探究被试主要通过面孔的哪些区域进行面孔的识别(即面孔识别诊断区域)，我们使用基于 Matlab 语言开发的 stat4CI 工具包对上述四张分类图像的噪音模式进行像素丛聚检验(clusters of pixels tests)(Chauvin et al., 2005)。根据 Dotsch 和 Todorov(2012)的检验方法，我们首先对噪音模式的像素值进行高斯滤波(高斯的标准差 =4 像素)使其平滑。其次，我们对面孔区域内平滑后的噪音模式的像素值进行 Z 转换。最后，对 Z 转化后的数据进行双尾的丛聚检验($Z_{crit} \geqslant |2.3|, p < 0.05$)以揭示显著的像素区域。

第二阶段 评估分类图像

第二阶段的任务 1 旨在考察被描述为可信的个体是否会与更有吸引力的面孔表征相关联。该任务包括男性和女性刺激两个试次，呈现顺序随机。每个试次左右并排呈现同性别的分类图像：一张来自可信组，另一张来自不可信组，面孔分类图像呈现的左右位置随机。被试需要选出更有吸引力的面孔。

任务 2 重复并拓展了任务 1 的评价维度，被试不仅需要选出更有吸引力的面孔，还需要选出更可信、更聪明、更友善、更多积极表情、更刻薄、更贪婪、更具攻击性和更具支配性的面孔。在每个试次中，一组面孔对以及一个评价维度的问题(如您认为下面两张面孔哪张

看起来更可信)同时呈现，被试按键选择。这个任务总共有 18 个试次(2 性别×9 组评价)，试次呈现顺序随机，每个试次中面孔呈现的左右位置也随机。

2.2 结果

2.2.1 可信度操纵的有效性

对可信度和活跃度评分分别进行 2 (可信水平：可信 vs 不可信)× 2 (目标面孔性别：男性 vs 女性) 的被试间方差分析。对于可信度评分，只有可信水平的主效应显著，$F(1, 151) = 452.64$, $p < 0.001$, $\eta^2 = 0.75$，被试对可信条件件($M = 7.30$, $SD = 1.81$)下的目标人物的信任度高于不可信条件件($M = 1.82$, $SD = 1.34$)。而对于活跃度的评分，只有目标面孔性别的主效应显著，$F(1, 151) = 6.70$, $p = 0.011$, $\eta^2 = 0.042$，被试认为男性目标人物($M = 4.25$, $SD = 1.63$)比女性目标人物($M = 3.60$, $SD = 1.46$)更活跃。可信度操纵影响了可信度评分而不影响与之无关的活跃度评分，结果证实了可信度操纵的有效性。

2.2.2 吸引力评价

可信和不可信条件下，女性和男性目标面孔的分类图像见图 3。无论是女性还是男性的分类图像，所有被试均选择可信条件下的分类图像更有吸引力，卡方检验显示它们的 $X^2(1) = 40$, $p < 0.001$。

2.2.3 多维特质评价

表 1 列出了针对不同特质，选择可信条件下的面孔分类图像的评分者人数和比例(共 50 人)以及卡方检验的统计结果。针对男女面孔各自的卡方检验显示：更多的评分者选择可信条件下的面孔分类图像更具有积极的特征(吸引力、聪明、可信、积极表情和友善)；相反，更多的评分者选择不可信条件下的面孔分类图像更具有消极的特征(刻薄、贪婪、攻击性和支配性)。我们进一步把性别纳入卡方检验(当某个组的样本数少于 5 时，显著性水平基于 Fisher 精确检验)，发现所有特质均不显著(见表 1 最后一列)，提示面孔性别并不影响被试的选择。

表 1 评价不同特质时选择可信条件下的面孔分类图像的评分者人数(比例)及卡方检验结果

特质	女性面孔 (实验 1)	X^2	男性面孔 (实验 2)	X^2	X^2(纳入性别)
吸引力	49 (98%)	46.08***	49 (98%)	46.08***	0
聪明	43 (86%)	25.92***	43 (86%)	25.92***	0

可信	49 (98%)	46.08^{***}	50 (100%)	50^{***}	1.01
积极表情	49 (98%)	46.08^{***}	47 (94%)	38.72^{***}	1.04
友善	49 (98%)	46.08^{***}	49 (98%)	46.08^{***}	0
刻薄	10 (20%)	18.00^{***}	4 (8%)	35.28^{***}	2.99
贪婪	6 (12%)	28.88^{***}	5 (10%)	32.00^{***}	0.10
攻击性	6 (12%)	28.88^{***}	8 (16%)	23.12^{***}	0.33
支配性	17 (34%)	5.12^{*}	16 (32%)	6.48^{*}	0.05

注：$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$，下同；卡方检验自由度均为 1

2.2.4 面孔识别诊断区域

像素丛聚检验的结果如图 4 所示。红色和绿色的区域代表显著的区域，即面孔识别的诊断区域。这些区域包括眼睛、鼻子、嘴巴以及少许的头发、脸颊和耳朵。其中，绿色的丛聚代表这块区域噪音的像素值越小时(会导致该分类图像的这块区域的颜色更暗)，被试更容易选择为目标面孔。相反，红色的丛聚代表这块区域噪音的像素值越大时(会导致该分类图像的这块区域的颜色更亮)，被试更容易选择为目标面孔。

可信条件 不可信条件

女性
目标面孔

男性
目标面孔

图 4 像素丛聚检验揭示的显著的丛聚

3 实验 2

实验 2 考察实验 1 中被描述为可信(或不可信)的个体面孔表征特征是否与可信(或不可信)的群体面孔表征特征有更多的相似性。为了产生可信/不可信群体的面孔表征(或者称可信

/不可信群体的面孔原型)，实验 2 的被试并没有观看个体面孔照片，也无需再认个体面孔，他们直接在添加了噪音的面孔中选出可信面孔。

3.1 研究方法

3.1.1 被试

招募了 20 名大学生被试(10 名女生)，平均年龄 19.95 岁，标准差 1.10 岁，视力或矫正视力正常。研究获得了浙江师范大学伦理委员会的批准。被试在参与实验前签署书面知情同意书，实验后获得一定的金钱报酬。

3.1.2 材料

反相关图像分类任务中的刺激与实验 1 相同。

3.1.3 实验流程

被试需要完成两个反相关图像分类任务(男性和女性面孔任务，任务顺序在被试间平衡)，以获得可信和不可信群体的面孔表征的分类图像(Dotsch & Todorov, 2012)。与实验 1 不同，被试无需执行印象形成任务，仅仅只需要完成反相关图像分类任务：从两张模糊面孔中选择一张与其脑海中和可信面孔最相似的面孔，总共完成 640 个试次。根据 Dotsch 和 Todorov (2012)的研究，被试选择的图像的噪音平均后叠加在面孔底片上即产生了可信群体的面孔表征，而未被选择的图像的噪音平均后叠加在面孔底片上即产生了不可信群体的面孔表征。

3.1.4 数据处理

为了探究不同面孔表征特征间的相似性，我们提取出两个实验产生的分类图像的面孔区域内的噪音模式的像素亮度值向量化后求皮尔逊相关(Dotsch & Todorov, 2012)。之所以没有直接对分类图像的像素值做相关，是为了避免面孔底片带来的冗余相关，因为所有同性别的分类图像均由相同的面孔底片叠加各自的噪音模式产生。此外，由于噪音模式的叠加能改变面孔底片的样貌，噪音模式可以认为是不同面孔表征独有的特征。为检验相关系数的显著性，使用 bootstrap 方法获得了上述相关系数的 95%置信区间。如果相关系数的 95%置信区间不包含 0，则可以认为相关系数与 0 显著不同。

为了比较两个相关系数之间的差异，我们使用了两种方法。第一种采用 Zou (2007)提出的方法，使用 R 软件(R Core Team, 2015)中的 cocor 包(Diedenhofen & Musch, 2015)计算两个相关系数之间差异的置信区间。如果 95%置信区间不包括 0，则可以认为这两个相关系数是显著不同的。第二种方法采用置换检验的方法(permutation test)，把真实的相关系数的差异值与随机置换产生的零分布做比较。为了产生零分布，我们随机打乱面孔区域内噪音像素值的

空间位置(比如第 100 个位置和第 200 个位置的像素值互换)以破坏原图像的结构并计算相关系数的差异。如此随机置换 1000 次，产生 1000 个相关系数差异值即形成零分布。最后，计算真实的相关系数差异值在零分布的位置即可计算 p 值。针对双尾检验，真实值在零分布的前 2.5%或者后 97.5%为显著。本研究重点关注的是，实验 1 中可信条件下的目标个体的面孔分类图像特征(噪音模式)是否会与可信群体面孔表征的分类图像特征(噪音模式)有更多的相似性，而不可信条件下的目标个体的面孔分类图像特征(噪音模式)是否会与不可信群体面孔表征的分类图像特征(噪音模式)有更多的相似性。

为了进一步控制相同面孔底片带来的系统偏差，我们使用偏相关的方法，把面孔底片的像素值向量化后纳入控制变量，再次求分类图像噪音模式之间的相关。两个相关系数之间的差异检验使用上述的置换检验的方法。

由于对图像像素值的向量化会丢失图像像素区块之间的空间联系，求向量化后的像素值的皮尔逊相关可能并不是最好的表征图像之间相似程度的方法。我们进一步使用 Matlab 软件里的 ssim 函数计算图像间的结构相似性指标(structural similarity index measure, SSIM)测量不同噪音模式间的相似性(Wang et al., 2004)。为了检验两个 SSIM 之间的差异是否达到统计显著性，我们依然使用了置换检验的方法，把真实的 SSIM 差异值与随机置换产生的零分布做比较计算 p 值。

3.2 结果

图 5 展示了可信和不可信群体面孔的分类图像。

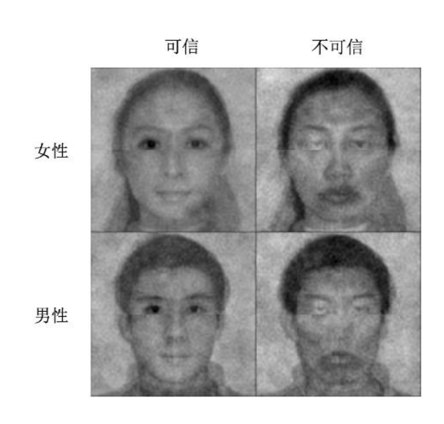


图 5 可信和不可信群体面孔的分类图像(所有被试的平均)

表 2 展示了不同分类图像的噪音模式之间的相关系数及对应的 95%置信区间。相关系数差异比较发现无论是男性还是女性面孔，可信条件下的目标面孔分类图像的噪音模式与可信群体面孔分类图像的噪音模式的相似性高于与不可信群体面孔分类图像的噪音模式的相

似性。此外，对于男性面孔，不可信条件下的目标面孔分类图像的噪音模式与不可信群体面孔分类图像的噪音模式的相似性高于与可信群体面孔分类图像的噪音模式的相似性。但是，对于女性面孔，不可信条件下的目标面孔分类图像的噪音模式与可信群体面孔分类图像的噪音模式的相似性高于与不可信群体面孔分类图像的噪音模式的相似性。

表 2 不同分类图像的噪音模式之间的相关系数及对应的 95%置信区间

		可信群体	不可信群体	相关系数差异	相关系数差异的置信区间
男性	可信个体	0.446 [0.498 0.507]	-0.170 [-0.182 -0.163]	0.616***	[0.610 0.623]
	不可信个体	0.028[0.020 0.036]	0.230[0.223 0.237]	-0.202***	[-0.210 -0.195]
女性	可信个体	0.514 [0.510 0.519]	-0.256 [-0.268 -0.247]	0.770***	[0.765 0.777]
	不可信个体	0.233[0.225 0.238]	0.048[0.042 0.058]	0.185***	[0.178 0.192]

注：相关系数的差异检验使用置换检验的方法，相关系数差异的置信区间使用 Zou (2007)的方法

当把面孔底片纳入控制变量后，不同分类图像的噪音模式之间的偏相关结果与上文的相关结果一致，见表 3。

表 3 在控制面孔底片后，不同分类图像的噪音模式之间的偏相关系数及对应的 95%置信区间

		可信群体	不可信群体	相关系数差异
男性	可信个体	0.443 [0.436 0.450]	-0.156 [-0.167 -0.147]	0.599***
	不可信个体	0.029[0.023 0.039]	0.230[0.224 0.237]	-0.201***
女性	可信个体	0.509 [0.505 0.513]	-0.271 [-0.281 -0.263]	0.780***
	不可信个体	0.238[0.230 0.243]	0.052[0.045 0.060]	0.186***

注：相关系数的差异检验使用置换检验的方法

表 4 展示了不同分类图像的噪音模式之间的 SSIM。差异比较发现无论是男性还是女性面孔，可信条件下的目标面孔分类图像的噪音模式与可信群体面孔分类图像的噪音模式的结构相似性高于与不可信群体面孔分类图像的噪音模式的结构相似性。此外，不可信条件下的目标面孔分类图像的噪音模式与不可信群体面孔分类图像的噪音模式的结构相似性高于与可信群体面孔分类图像的噪音模式的结构相似性。

表 4 不同分类图像的噪音模式之间的结构相似性(SSIM)

		可信群体	不可信群体	结构相似性差异
男性	可信个体	0.383	0.122	0.261***
	不可信个体	0.244	0.265	-0.021***
女性	可信个体	0.391	0.062	0.329***
	不可信个体	0.187	0.279	-0.092***

注：结构相似性的差异检验使用置换检验的方法

4 讨论

本研究使用反相关图像分类技术探究了感知到的信任程度对个体面孔心理表征的影响及其潜在机制。实验 1 发现同一张目标面孔被赋予不同的可信度描述时，被认为可信(相对于不可信)人物的面孔表征，也被认为更具吸引力。这个结果说明特质推断对面孔吸引力的感知不局限于自我报告的评分，还包括面孔表征的形成等知觉过程。以往研究发现对社会群体(如男性群体)的表征不仅会受到自下而上的视觉特征(如较大的下巴)的影响，还会受到自上而下的刻板印象(如攻击性)的影响(Bagnis et al., 2019; Freeman & Ambady, 2011; Freeman & Johnson, 2016; Hehman et al., 2014)。本研究结果与前人一致，说明“人们在脑海中所表征的面孔样貌并不是他们实际看到的”，自上而下的因素(如对人格特质的了解)也可以影响人们对他人面孔的表征。

实验 1 还表明，人们不仅觉得可信个体的面孔表征更有吸引力，而且还会被描述为具有更多其他的积极特质。以往研究已经证实可信度和吸引力之间存在紧密联系，表明对他人的信任程度依赖于他人面孔的吸引力(Gutierrez-Garcia et al., 2019; Oosterhof & Todorov, 2008;

Willis & Todorov, 2006; Xu et al., 2012)。神经影像学的研究也表明，对面孔可信度和吸引力的判断涉及相似的脑区并且这种感知是自发的(Bzdok et al., 2011; Engell et al., 2007)。本研究的结果与之前的研究一致。此外，可信度是一个多维度的特质，包括如诚实、可靠和仁慈等不同方面(Mayer et al., 1995)。本研究进一步表明，信任的印象也会影响对个体其他积极特征的感知(如友善、聪明等)。

实验 1 还进一步探索了面孔识别时的诊断区域，这些区域主要包括眼睛、鼻子和嘴巴，以及少许的头发以及面孔轮廓(脸颊和耳朵边缘)(图 4)。这个结果与 Dotsch 和 Todorov(2012)的研究很一致，尽管他们让被试对面孔进行可信和支配(dominance)程度的判断。这些结果说明人们在进行不同的社会判断时，往往从面孔相似的区域提取信息，同时特别依赖眼鼻嘴这些关键的面孔区域。此外，诊断区域的揭示也突出了反相关图像分类技术能提供非常丰富的数据。

实验 2 在实验 1 基础上探索了感知信任影响心理面孔表征的潜在机制。本研究假设，当个体在形成目标面孔的心理表征时，会对某个社会群体的刻板印象整合到该目标个体中。具体到本研究中则表现为，对可信(或不可信)的描述会导致个体将目标人物划分至可信(或不可信)的群体。然后，人们会将脑海中可信(或不可信)的群体刻板面部特征融入到目标人物的面部特征中。在此过程下，感知到的信任水平会使感知者对目标人物的客观面孔物理特征的表征产生偏差。实验 2 结果支持了这一假设：结构相似性指标揭示无论目标人物是男性还是女性，可信(或不可信)的目标人物的面孔表征特征与可信(或不可信)群体的面孔表征特征有更多的相似性。但是，基于皮尔逊相关/偏相关的结果却发现当目标人物为女性时，不可信目标人物的面孔表征特征也与可信群体的面孔表征特征有更多的相似性。这可能是因为对图像像素值向量化后求相关会丢失图像像素区块之间的空间联系，导致它不能很好地表征图像之间的相似程度。此外，人们很难去丑化一个人，表 2 和表 3 揭示的不可信目标人物和不可信群体之间的相似性要小于可信目标人物和可信群体之间的相似性印证了这一点。人们也会认为女性要比男性更可信(Buchan et al., 2008; Dong et al., 2018)，因此，在表征的过程中人们可能也会把更多的可信特征添加到不可信女性目标面孔上，这也可能是导致女性结果并不稳定的原因。未来研究需要进一步探究这些可能性。

未来研究可以探究形成的面孔表征是否会对外显行为产生影响。比如，在与他人进行首次面对面的社交互动时所形成的面孔表征是否会对后续的非面对面社会互动(如电话或电子邮件)中做出的商业和政治决策等产生影响。另外，还可以探究塑造个体面孔心理表征的可

能性。许多研究表明记忆可以被重建(Barrouillet & Camos, 2014; Loftus, 1975; Mecklinger et al., 2016)，心理面孔表征作为记忆的一种表现形式也应该可以被重构。未来研究可以探索这种重构的可能性，以及重构后个体对该目标人物的外显行为或态度是否也会改变。从更宏观的视角来讲，后续研究可以考虑开发有效的干预手段以重建对其他种族面孔的心理表征，并考察面孔表征的重构能否有效促进种族间的积极互动。

本研究也有一些局限性，值得未来研究进一步探索。首先，使用反相关图像分类技术获得的面孔分类图像只能算是真实心理表征的近似值，因为面孔底片以及噪音的选择会影响面孔表征的样貌(Dotsch & Todorov, 2012)。毫无疑问，如果换一张面孔底片做出来的分类图像会与当前研究得到的不一样。相比于面孔底片，噪音的选择对面孔表征样貌的影响要小很多。Dotsch 和 Todorov(2012)的研究使用了两组随机产生的噪音(每组都是 300 对正负噪音)，发现结果很一致。这可能是因为反相关分类图像技术往往需要较多的试次进行平均(文献中一般都是多于 300 试次，本研究用了 640 个试次)，这降低了噪音选择的影响。即使有这些混杂因素的影响，我们也必须强调采用反相关图像分类技术做的研究往往关注的是类别间的相对差异(比如这一类别的面孔表征是不是比另一类别面孔表征更有吸引力)，而不是某类面孔表征的绝对样貌。本研究主要关注的是可信/不可信操纵造成的个体面孔表征的差异或是可信/不可信个体与可信/不可信群体面孔表征相似性的差异。虽然同性别面孔的底片是一样的，叠加的噪音也是一样的，但最终差异还是显著的，说明心理表征的变化主要来源于对于目标面孔可信度的描述不同。但是无论如何，将混杂的因素分离并得到真实的心理表征是值得未来研究进一步探索的。其次，本研究主要探究个体面孔表征受到可信/不可信描述的影响，但又不可能把所有个体研究一遍，因此本研究的结果是特异于实验 1 选择的两个个体还是对于大部分个体都成立需要未来研究进一步拓展和验证。特别是本研究选取了具有中等吸引力面孔的个体，因此能不能推广到本身吸引力较高和较低的面孔上值得一探究竟。

总之，本研究首次发现感知到的信任程度会影响个体面孔表征，一个可信的个体的面孔会被赋予更多积极的特征。因此，灰姑娘能变成白雪公主：即使一个人的长相一般，只要拥有美好的心灵，人们心目中的这个人的相貌也会更具吸引力。

参考文献

Bagnis, A., Celeghin, A., Mosso, C. O., & Tamietto, M. (2019). Toward an integrative science of social vision in intergroup bias. *Neuroscience and Biobehavioral Reviews, 102*, 318–326.

Barrouillet, P., & Camos, V. (2014). *Working memory: Loss and reconstruction*. Psychology Press.

Bascandziev, I., & Harris, P. L. (2014). In beauty we trust: Children prefer information from more attractive informants. *British Journal of Developmental Psychology, 32*(1), 94–99.

Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion, 8*(4), 479–493.

Bonnefon, J. F., Hopfensitz, A., & De Neys, W. (2017). Can we detect cooperators by looking at their face? *Current Directions in Psychological Science, 26*(3), 276–281.

Buchan, N. R., Croson, R. T. A., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the Investment Game. *Journal of Economic Behavior & Organization*, *68*(3), 466–476.

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., Laird, A., & Eickhoff, S. B. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure & Function, 215*(3-4), 209–223.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision, 5*, 659–667.

Chen, F. F., Jing, Y., & Lee, J. M. (2014). The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology, 51*, 27–33.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *Plos One, 10*(4), 12.

Dion, K., Walster, E., & Berscheid, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*(3), 285.

Dion, K. K. (1972). Physical attractiveness and evaluation of children's transgressions. *Journal of Personality and Social Psychology, 24*(2), 207–213.

Dong, Y., Liu, Y., Jia, Y., Li, Y., & Li, C. (2018). Effects of facial expression and facial gender on judgment of trustworthiness: The modulating effect of cooperative and Competitive settings. *Frontiers in Psychology*, *9*, e2022.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*(5), 562–571.

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science, 19*(10), 978–980.

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology, 100*(6), 999–1014.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19*(9), 1508–1519.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review, 118*(2), 247–279.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences, 20*(5), 362–374.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science, 14*(5), 505–509.

Gutierrez-Garcia, A., Beltran, D., & Calvo, M. G. (2019). Facial attractiveness impressions precede trustworthiness inferences: Lower detection thresholds and faster decision latencies. *Cognition & Emotion, 33*(2), 378–385.

Hehman, E., Ingbretsen, Z. A., & Freeman, J. B. (2014). The neural basis of stereotypic impact on multiple social categorization. *Neuroimage, 101*, 704–711.

Karremans, J. C., Dotsch, R., & Corneille, O. (2011). Romantic relationship status biases memory of faces of attractive opposite-sex others: Evidence from a reverse-correlation paradigm. *Cognition, 121*(3), 422–426.

Krosch, A. R., & Amodio, D. M. (2014). Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences of the United States of America, 111*(25), 9079–9084.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin, 126*(3), 390–423.

Lin, C. J., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of english trait words. *Nature Communications, 12*(1), 15, 5168.

Lloyd, E. P., Sim, M., Smalley, E., Bernstein, M. J., & Hugenberg, K. (2020). Good cop, bad cop: Race-based differences in mental representations of police. *Personality and Social Psychology Bulletin, 46*(8), 1205–1218.

Loftus, E. F. (1975). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal, 15*(3), 6.

Maister, L., De Beukelaer, S., Longo, M. R., & Tsakiris, M. (2021). The self in the mind's eye: Revealing how we truly see ourselves through reverse correlation. *Psychological Science, 32*(12), 1965–1978.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

Mecklinger, A., Rosburg, T., & Johansson, M. (2016). Reconstructing the past: The late posterior negativity (LPN) in episodic memory studies. *Neuroscience and Biobehavioral Reviews, 68*, 621–638.

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience, 8*(3), 285–299.

Moon, K., Kim, S., Kim, J., Kim, H., & Ko, Y. G. (2020). The mirror of mind: Visualizing mental representations of self through reverse correlation. *Frontiers in Psychology, 11*, 8.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087–11092.

Paunonen, S. V. (2006). You are honest, therefore I like you and find you attractive. *Journal of Research in Personality, 40*(3), 237–249.

R Core Team. (2015). http://www.R-project.org/

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology, 106*(6), 897–911.

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology, 113*(4), 1056–1078.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*, 519–545.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813–833.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology, 83*(5), 1051–1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39*(6), 549–562.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology, 87*(4), 482–493.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility

to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592–598.

Xu, F., Wu, D. C., Toriyama, R., Ma, F. L., Itakura, S., & Lee, K. (2012). Similarities and differences in chinese and caucasian adults' use of facial cues for trustworthiness judgments. *Plos One, 7*(4), 9.

Young, A. I., Ratner, K. G., & Fazio, R. H. (2014). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science, 25*(2), 503–510.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12*(4), 399–413.

Can Cinderella become Snow White? The Influence of Perceived Trustworthiness on the Mental Representation of Faces

LI Qinggong[1], FANG Wei[2,3], HU Chao[4], SHI Dejun[5],

HU Xiaoqing[6], FU Genyue[7], WANG Qiandong[8]

(*[1] Zhejiang Philosophy and Social Science Laboratory for the Mental Health and Crisis Intervention of Children and Adolescents, College of Psychology, Zhejiang Normal University, Jinhua 321004, China*)

(*[2] Department of Psychology, Zhejiang Sci-Tech University, Hangzhou 310018, China*)

(*[3] Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton L8S 4L8, Canada*)

(*[4] Department of Medical Humanities, School of Humanities, Southeast University, Nanjing 211189, China*)

(*[5] School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China*)

(*[6] The State Key Laboratory of Brain and Cognitive Sciences, Department of Psychology, The University of Hong Kong, Hong Kong 999077, China*)

(*[7] Zhejiang Philosophy and Social Science Laboratory for Research in Early Development and Childcare, Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments, Department of Psychology, Hangzhou Normal University, Hangzhou 311121, China*)

(*[8] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University, Beijing 100875, China*)

Abstract

People often infer others' social traits, such as trustworthiness, from a glance at their face. Whereas previous studies have focused on how different facial cues influence social perception, the present study examined whether perception of a person's trustworthiness could influence mental representations of that person's face, as well as the mechanisms underlying this process.

Two experiments were conducted. Experiment 1 was designed to test whether a target person described as trustworthy would be represented in the perceiver's mind as more attractive than the same person described as untrustworthy. Participants were instructed to form an impression about a target person's trustworthiness by viewing the person's face paired with a description labeling them as trustworthy or untrustworthy. The reverse correlation image classification (RCIC) technique was then used to visualize the participants' mental representations of the target person's

face. A separate group of participants was recruited to evaluate the attractiveness and other traits of the generated mental representation images. Experiment 2 aimed to determine a possible underlying mechanism by exploring whether the mental representations of the trustworthy (or untrustworthy) target persons' faces in Experiment 1 shared more similarities with those of the trustworthy (or untrustworthy) faces at a group level (i.e., prototypes of trustworthy or untrustworthy faces). To achieve this goal, we recruited participants to complete an alternate RCIC task in which they selected which of two faces appeared more trustworthy, producing mental representation images for trustworthy and untrustworthy faces at a group level. The features of these prototypical trustworthy and untrustworthy faces were then compared with those of the target person from Experiment 1.

In Experiment 1, mental representations of a face described as trustworthy were found to be more attractive than those of the same face described as untrustworthy. Furthermore, raters attributed additional desirable traits, such as friendly, intelligent, and positive, to the representation of the trustworthy person. In Experiment 2, we found that the mental representation of the face labeled as trustworthy in Experiment 1 shared more similarities with the prototypical trustworthy face produced in Experiment 2 than with the prototypical untrustworthy face.

In sum, our findings suggest that the perception of a person's trustworthiness can influence mental representations of that person's face. When people perceive an individual as trustworthy (or untrustworthy), they may superimpose the corresponding schema features in their minds onto the physical characteristics of the perceived individual's face, leading to a reconfiguration of the face representation. Our study underscores the importance of top-down factors in shaping face representations.

Keywords person perception, reverse correlation image classification technology, mental representation, attractiveness, trustworthiness